A PERCEPTUAL APPROACH FOR SEMANTIC IMAGE RETRIEVAL

Dejan Depalov, Thrasyvoulos Pappas

EECS Department, Northwestern University
2145 Sheridan Rd., Evanston, IL 60208
{depalov,pappas}@ece.northwestern.edu

Dongge Li, Bhavan Gandhi

Motorola Labs, Motorola
1303 E. Algonquin Rd., Schaumburg, IL 60196
{dongge.li,bhavan.gandhi}@motorola.com

ABSTRACT

The rapid growth of digital imaging technology and the accumulation of large collections of digital images has created the need for efficient and intelligent schemes for image classification and retrieval. Since humans are the ultimate users of most retrieval systems, it is important to organize the contents semantically, according to meaningful categories. We propose a novel approach for assigning semantic labels to image segments, which together segment layout information can lead to content-based image classification and retrieval. The proposed approach relies on a perceptually based, spatially adaptive, color-texture segmentation scheme. We derive segment-wide features (color and spatial texture). These features serve as medium level descriptors that can effectively bridge the "semantic gap" between low and high level descriptors. The segment classification into semantic categories is based on linear discriminant analysis techniques. We demonstrate the effectiveness of the proposed approach on a database that includes 5000 segments from approximately 2000 photographs of natural scenes.

1. INTRODUCTION

The goal of content-based image retrieval (CBIR) is to facilitate the automatic indexing of large image repositories based on image semantics. The field of CBIR has been extensively researched in recent years with main emphasis on query by example based on matching low-level image features, such as color and texture, with or without relevance feedback by the user. (See [1] for a review.) However, none of the proposed approaches has achieved satisfactory performance because it has been difficult to infer semantic meaning from low-level features. This gap between low-level image features and high level semantics is known as the *semantic gap.*

Recently several approaches have been proposed that attempt to bridge the semantic gap in order to produce an automated CBIR system. Most of the approaches utilize some kind of image segmentation scheme, to extract the image regions and then try to obtain their content as well as their context within an image. Zhu *et al.* [2] partition the image into equal size blocks and index the regions using a codebook whose entries are obtained from the block features. Wang *et al.* [3] also propose a codebook based approach, whereby the codebook is used to segment the image based on the statistics of the region color and texture features. Their approach also attempts to take into account properties of the neighboring regions. Pan *et al.* [4] use a simple segmentation technique to segment an image into regions and extract their features. Each region is is given a label called a blob-token. The authors attempt to find the association among the blob-tokens and the associated captions to index the image. Li and Wang [5] use a statistical modeling approach in which images of a given concept are regarded as the instances of a random process characterizing this concept. Their method utilizes 2D hidden Markov models to calculate a measure of association between the image and the textual description of a concept. Finally, Mojsilovic and Rogowitz [6] attempt to link low-level image features directly to image semantics.

Despite all this effort, the effectiveness of CBIR systems has not been satisfactory and they are still a long way from matching the performance of the human visual system (HVS). This paper proposes a novel approach for image indexing that utilizes perceptual models for image segmentation and classification. Our group has developed an adaptive, perceptual color-texture segmentation algorithm that combines knowledge of human perception with an understanding of signal characteristics to segment natural scenes into perceptually, semantically uniform regions [7]. We describe how this new segmentation methodology can be used for image labeling and classification. This requires the derivation of region-wide color and texture features. Such medium level descriptors are the key to bridging the gap between low-level image primitives and high-level image semantics. However, these descriptors are meaningful only if the segments are perceptually/semantically relevant. Thus, the success of the proposed approach is critically dependent on the segmentation methodology proposed in [7]. We present techniques for assigning labels to image segments based on color and spatial texture descriptors. Further improvements can be achieved by incorporating the segment location, size, and boundary shape, as well as the properties of the neighboring segments. However, in this paper we focus on color and spatial texture. We demonstrate the effectiveness of the proposed approach using a database of approximately 2000 images of natural scenes, which were segmented using the algorithm in [7]. Our results indicate that the proposed approach can offer significant performance improvements over existing approaches.

The focus of this paper is on still images. The techniques we discuss, however, can also form the basis for content-based analysis of video sequences. We consider the domain of photographic images with a wide range of content (indoor and outdoor natural and manmade scenes).

2. COLOR-TEXTURE FEATURE SELECTION

We now review the color-texture features that were developed for the adaptive perceptual segmentation algorithm proposed in [7]. These features can also be used for segment classification.

The segmentation approach [7] incorporates models of human

This work was supported by the National Science Foundation (NSF) under Grant No.CCR-0209006. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF. This work was also supported by the Motorola Center for Telecommunications at Northwestern University.

(a) (b) (c) (d)

Fig. 1. Color-texture image features and segmentation. (a) original color image (b) adaptive dominant colors (c) texture classes (smooth regions are shown in black, horizontal in gray, and complex in white) (d) final segmentation

perception and signal characteristics. It is based on two types of spatially adaptive features. The first provides a localized description of the color composition of the texture and the second models the spatial characteristics of its grayscale component.

The color composition feature exploits the fact that the HVS cannot simultaneously perceive a large number of colors. In addition, it accounts for the spatially varying image characteristics and the adaptive nature of the HVS. It thus consists of a small number of spatially adaptive dominant colors and the corresponding percent occurrence of each color in the vicinity of a pixel:

$$f_c(x, y, N_{x,y}) = \{(c_i, p_i), i = 1, \ldots, M, p_i \in [0, 1]\} \quad (1)$$

where c_i is a 3-D color vector and p_i is the corresponding percentage. $N_{x,y}$ denotes the neighborhood of the pixel at (x, y) and M is the number of dominant colors in $N_{x,y}$; a typical value is $M = 4$. The spatially adaptive dominant colors are obtained using the adaptive clustering algorithm (ACA) [8]. An example is shown in Figure 1(b). The perceptual similarity between two color composition feature vectors is based on the "Optimal Color Composition Distance (OCCD)," which finds the optimal mapping between the color composition features of two segments and computes the average distance between them in the *CIE L*a*b** color space.

The spatial texture feature extraction is based on a multiscale frequency decomposition with four orientation subbands (horizontal, vertical, +45°, -45°). Here, we use a one-level steerable filter decomposition with four orientation subbands. The local energy of the subband coefficients is used as a simple but effective characterization of spatial texture. At each pixel location, the maximum of the four subband coefficients determines the texture orientation. A median filtering operation boosts the response to texture within uniform regions and suppresses the response resulting from to transitions between regions. Pixels are then classified into smooth and non-smooth classes, and non-smooth pixels are further classified on the basis of dominant orientation, as horizontal, vertical, +45°, -45°, and complex (i.e., no dominant orientation). An example is shown in Figure 1(c).

The segmentation algorithm combines the color composition and spatial-texture features to obtain segments of uniform texture. It is a fairly elaborate algorithm that relies on spatial texture to determine the major structural composition of the image and combines it with color, first to estimate the major segments, and then to obtain accurate and precise localization of the border between regions.

Several critical parameters of the texture features and segmentation algorithm can be determined by subjective tests [9]. These include thresholds for the smooth/non-smooth classification, for determining the dominant orientation, and for the color-composition feature similarity. The goal of the tests is to relate human perception of isolated (context-free) texture patches to the statistics of natural textures. Experimental results demonstrate that this perceptual tuning leads to significant improvements in segmentation performance.

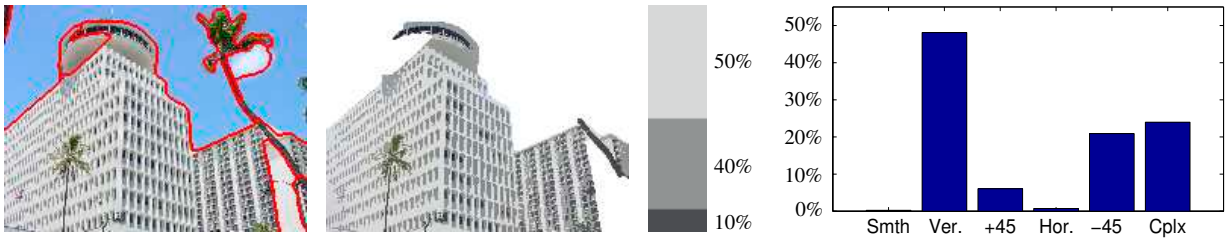
3. SEGMENT-WIDE FEATURE EXTRACTION

We now discuss the development of medium level color and spatial texture descriptors. While image segmentation requires a combination of local and global features [7], region classification requires segment-wide features. Thus, for each segment, we recalculate the color composition and spatial texture features using only information from within the segment, that is, the local averages and medians are computed across and strictly within the segment. The texture features of the segment can be similarly described by the percentage of smooth, horizontal, vertical, +45°, -45°, and complex pixels . An example is shown in Fig. 2, where (a) shows a segmented image, (b) shows a selected segment, (c) shows the segment-wide color composition (dominant colors and percentages), and (d) shows the region-wide spatial texture features (percentage of smooth, horizontal, vertical, +45°, -45°, and complex pixels).

Observe that there is an asymmetry between the two types of features. The spatial texture features consist of six labels and the corresponding percentages, while the color composition features consist of up to four dominant colors (which take essentially a continuum of values) and the associated percentages. In order to reduce the dimensionality of the color composition features, we assign color names to the dominant colors of each region. The procedure for assigning color names can be found in [10]. The selected color names (labels) are consistent with a *National Bureau of Standards* recommendation for color names. The syntax contains color names for 267 regions in color space, and employs English terms to describe colors along the three dimensions of the color space: hue, lightness and saturation. There are seven discrete values for lightness, five discrete values for saturation, and a basic set of eleven prototypical hues, as shown in Table 1. Thus, if we assign labels based on hue only, we end up with 14 labels (and corresponding percentages) instead of a continuum of color values, which establishes a symmetry with the spatial texture features. The use of a limited number of colors is consistent with Boynton's study, which found that when people are asked to categorize colors, the number of perceptually distinguishable color categories is small. (See his 1989 paper "Eleven colors which are almost never confused" [11].)

4. SEMANTIC LABELING

Once the medium level descriptors have been identified, the task is to extract semantic labels, first at the segment level and then for the entire image. Recent subjective experiments have identified important semantic categories that people use for image organization and retrieval [6]. Two important dimensions in human similarity perception are "natural" versus "man-made," and "human" versus "non-human." It was also found that certain cues, such as "sky," "water," "mountains," etc., have an important influence in human image perception. Rather than trying to obtain a complete and detailed de-



(a) (b) (c) (d)

Fig. 2. Segment-wide feature extraction. (a) Segmented image. (b) Selected segment. (c) Its color composition. (d) Its texture composition.

Hue Primary	Hue Secondary	Saturation	Lightness	Achromatic
Red	Reddish	Grayish	Blackish	Black
Orange	Brownish	Moderate	Very-dark	Gray
Brown	Yellowish	Medium	Dark	White
Yellow	Greenish	Strong	Medium	
Green	Bluish	Vivid	Light	
Blue	Purplish		Very-light	
Purple	Pinkish		Whitish	
Pink				
Beige				
Magenta				
Olive				

Table 1. Color Naming Syntax

scription of every object in the scene, then, this information suggests that for image classification, it may be sufficient to isolate segments of such perceptual significance.

Our first goal is to assign labels to image segments. To this end, we have assembled a vocabulary of labels consistent with the above findings, as well as those used in annotation of the NIST TRECVID 2003 development set [12]. The set of labels we selected is a subset of NIST lexicon. To describe the content of an image we use two types of labels, **segment** and **scene** labels. The segment labels describe the semantics of a particular segment (e.g., building, sky), while the scene labels describe the (higher-level) semantic content of the image (e.g., beach scene). The latter cannot be inferred from a particular image segment alone. The segment labels we chose are shown in Table 2, and are arranged in a hierarchical manner, at the top of which are the natural, man-made, human categories, and only leaf nodes are used in the annotation.

Learning and Classification

To demonstrate the effectiveness of the proposed approach, we conducted a set of simple experiments with a database of approximately 2000 photographs. The majority of the images were obtained from the Corel Stock Photo Library. Additional images were obtained from a Key Photos Library and the investigator's personal repository. The images in the database cover a variety of outdoor scenes, with a wide range of themes. As the initial focus of our experiments was on natural vs. man-made classification, we did not include any scenes with humans or animals. The human detection problem (especially face detection) is well-studied in the literature [13], and the existing techniques can easily be combined with the proposed approach. The problem of animal detection is more complicated because of natural camouflage. However, we believe that the proposed approach is capable of segmenting and detecting animals.

The images were segmented using the adaptive perceptual color-texture image segmentation algorithm [7] described above, and the resulting segments were manually labeled to be used as the ground truth for supervised learning and testing. Each segment was assigned exactly one label. Segments whose area was less than two percent of total image area were not considered. This resulted in approximately 5000 labeled segments, 80% of which were used for training and the rest for testing.

For training and classification we have considered several methods including: linear discriminant analysis (LDA) [14], K-nearest neighbors (KNN), support vector machines (SVM) [15,16], and Gaussian mixture models (GMM) [17, 18]. We did not consider the KNN approach due to its inability to compactly represent the classification function. SVM methods usually perform well, but their performance is dependent on good selection of the kernel function. Since the kernel function is usually chosen heuristically, there is no guarantee that the same kernel function will perform well if new classes are introduced, or if a classification is performed deeper in the hierarchy. This leaves us with a choice between GMM and LDA. Our experiment indicate that LDA significantly outperforms GMM. This is because our feature vector clusters are not Gaussians, and a small number of Gaussians is not enough to approximate their distribution. Increasing the number of Gaussians requires increasing the training set and computational complexity. Furthermore, the expectation maximization method (EM) used to build a GMM converges only locally.

The LDA was applied hierarchically, starting from the top node down to the leafs, obtaining a finer classification at each step. In the experiment described here we used only spatial texture and color-name composition descriptors. We mapped the dominant colors onto the 11 prototypical hues shown in Table 1 and the three achromatic colors (black, gray, and white) for a total of 14 colors. Thus, the feature vector contains a percentage for each of these 14 colors. The overall dimension of the feature vector is 20 (six textures and 14 colors). For the example in Fig. 2, the feature vector for the selected segment contains the percentages for the six texture categories, while all the color entries are zero except for gray (50%) and white (50%). The resulting segment label is "building/house" (man-made).

5. CLASSIFICATION RESULTS

We evaluated the performance of the proposed method using the standard measures that are used for search strategies in the literature. The **recall** is the ratio of the correctly labeled segments to the total number of relevant segments in the database (i.e., those with the particular label). The **precision** is the ratio of the correctly labeled segments to the total number of segments that the algorithm assigned to the particular label (both correct and incorrect). Both performance measures are expressed as percentages.

Our segment classification results are shown in Fig. 3, and compare favorably to the methods described in the literature (e.g., [19–

Natural				Man-made	Human	Animal
Vegetation	Sky	Landform	Water			
Grass Trees/bushes Forest Flowers	Day-sky Night-sky Sun Clouds Sunrise/Sunset	Snow Mountain/Hill Ground Pavement/Road[1]		Building/House Bridge Car Boat Airplane Other	Face Person People	

Table 2. Segment Labels

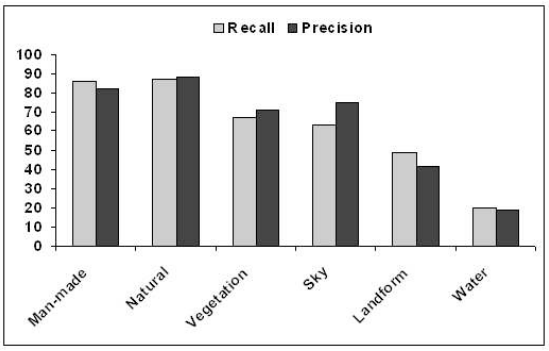


Fig. 3. Classification Results

24]). The precision and recall rates for the categories at the top of the hierarchy (natural vs. man-made) are quite impressive, especially if one takes into account the fact that we did not use the segment size, location, shape, and any information about neighboring segments. The rates for the categories further down in the hierarchy are not as impressive, but still compare well to the literature.

The poor performance for the water category can be attributed to inadequate feature selection. Since the color quntization is very coarse, water (whose primary hue is blue or green) gets confused with sky and sometimes vegetation or mountains. Our experiments indicate that expanding our color descriptors (e.g. to include lightness) and incorporating the location information in the feature vector of each segment can eliminate this confusion.

Overall, we believe that the inclusion of additional features (more color names, segment area, position, and shape), as well as the properties of the neighboring segments, will further improve performance. Finally, we also plan to use probabilistic layout models in order to combine the information from the segments in order to obtain an overall scene interpretation.

6. REFERENCES

[1] W.M. Smeulders, *et al.*, "Content-based image retrieval at the end of the early years," *IEEE Tr. PAMI*, vol. 22, pp. 1349–1379, Dec. 2000.

[2] L. Zhu, *et al.*, "Keyblock: An approach for content-based image retrieval," *ACM Multimedia 2000*, pp. 157–166, Oct. 2000.

[3] W. Wang, Y. Song, A. Zhang, "Semantics retrieval by content and context of image regions," *Proc. 15th Int. Conf. Vision Interface*, pp. 17–24, May 2002.

[4] J.Y. Pan, H.J. Yang, P. Duygulu, C. Faloutsos, "Automatic image captioning," *ICME*, 2004.

[1]While "pavement/road" is man-made, its features are almost identical to those for "ground."

[5] J. Li, J. Wang, "Automatic linguistic indexing of pictures by a atatistical modeling approach," *IEEE Tr. PAMI*, vol. 25, Sept. 2003.

[6] A. Mojsilović, B.E. Rogowitz, "Semantic metric for image library exploration," *IEEE Tr. MM*, vol. 6, pp. 828–838, Dec. 2004.

[7] J. Chen, *et al.*, "Adaptive perceptual color-texture image segmentation," *IEEE Tr. Im. Proc.*, vol. 14, pp. 1524–1536, Oct. 2005.

[8] T.N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Tr. Signal Proc.*, vol. 40, pp. 901–914, Apr. 1992.

[9] J. Chen, T.N. Pappas, "Experimental determination of visual color and texture statistics for image segmentation," Proc. SPIE Vol. 5666, pp. 227–236, Jan. 2005.

[10] A. Mojsilović, "A computational model for color naming and describing color composition of images," *IEEE Tr. Im. Proc.*, vol. 14, pp. 690–699, May 2005.

[11] R.M. Boynton, "Eleven colors that are almost never confused," Proc. SPIE Vol. 1077, pp. 322–332, Jan. 1989.

[12] C.Y. Lin, B.L. Tseng, J.R. Smith, "Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets." NIST TRECVID, 2003.

[13] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comp. Vision*, vol. 57, pp. 137–154, 2004.

[14] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*. Wiley, 2001.

[15] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[16] B. Schölkopf, C.J.C. Burges, A.J. Smola, eds., *Advances in Kernel Methods: Support Vector Learning*. MIT Press, 1998.

[17] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[18] B. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[19] J. Pan, H.J. Yang, P. Duygulu, C. Faloutsos, "Automatic image captioning," *ICME*, June 2004.

[20] K. Barnard, *et al.*, "Matching words and pictures," *J. Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[21] J. Li and J.Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Tr. PAMI*, vol. 25, pp. 1075–1088, 2003.

[22] J. Jeon, V. Lavrenko, R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," *SIGIR*, pp. 119–126, ACM, 2003.

[23] J.Z. Wang, J. Li, G. Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," *IEEE Tr. PAMI*, vol. 23, pp. 947–963, Sept. 2001.

[24] C. Carson, *et al.*, "Blobworld: A system for region-based image indexing and retrieval," *Proc. Third Int. Conf. Visual Information Systems*, pp. 509–516, Springer, June 1999.